Competition 2

Рысьмятова Анастасия

ВМК МГУ 317 группа

13.05.2015

Содержание

Похожий конкурс

Completed • $680 • 120 teams

Greek Media Monitoring Multilabel Classification (WISE 2014)

Mon 2 Jun 2014 – Tue 15 Jul 2014 (10 months ago)

- Использовать линейные классификаторы
- Использовать KNN
- Использовать регрессию и отсекать по порогу
- Смешивать различные решения

Проблемы

- Данные хранятся в spars матрице
- Большинство алгоритмов долго обучается на данных, поэтому сложно их настроить

Содержание

Первое решение

модель - *PassiveAggressiveRegressor*

$$\ell(\mathbf{w};(\mathbf{x},y)) = \begin{cases} 0 & y(\mathbf{w}\cdot\mathbf{x}) \geq 1 \\ 1 - y(\mathbf{w}\cdot\mathbf{x}) & \text{otherwise} \end{cases}.$$

INPUT: aggressiveness parameter $C > 0$
INITIALIZE: $\mathbf{w}_1 = (0,\ldots,0)$
For $t = 1,2,\ldots$

- receive instance: $\mathbf{x}_t \in \mathbb{R}^n$
- predict: $\hat{y}_t = \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t)$
- receive correct label: $y_t \in \{-1,+1\}$
- suffer loss: $\ell_t = \max\{0\ ,\ 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)\}$
- update:

1. set:

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2} \quad \text{(PA)}$$

$$\tau_t = \min\left\{C\ ,\ \frac{\ell_t}{\|\mathbf{x}_t\|^2}\right\} \quad \text{(PA-I)}$$

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}} \quad \text{(PA-II)}$$

2. update: $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$

Figure 1: Three variants of the Passive-Aggressive algorithm for binary classification.

Все параметры настраивались с помощью кросс-валидации использовалось отсечение по порогу с константой около 0.5 (константа настраивалась с помощью кросс-валидации)

Идеи улучшающие решение

- - Добавить KNN с метрикой 'cosine'
- - Добавить логистическую регрессию
- - Добавить Ridge регрессию
- - Добавить naive_bayes.BernoulliNB

Все это улучшало в 3 знаке после запятой.
В итоге результат на лидерборде - 0.51926

Random Forest

Random Forest - долго обучался но давал очень хороший результат на 100 деревьях качество около 0.51
Финальное решение использовало смесь линейных классификаторов с Random Forest почти с одинаковыми весами.